Introduction to Dataverse

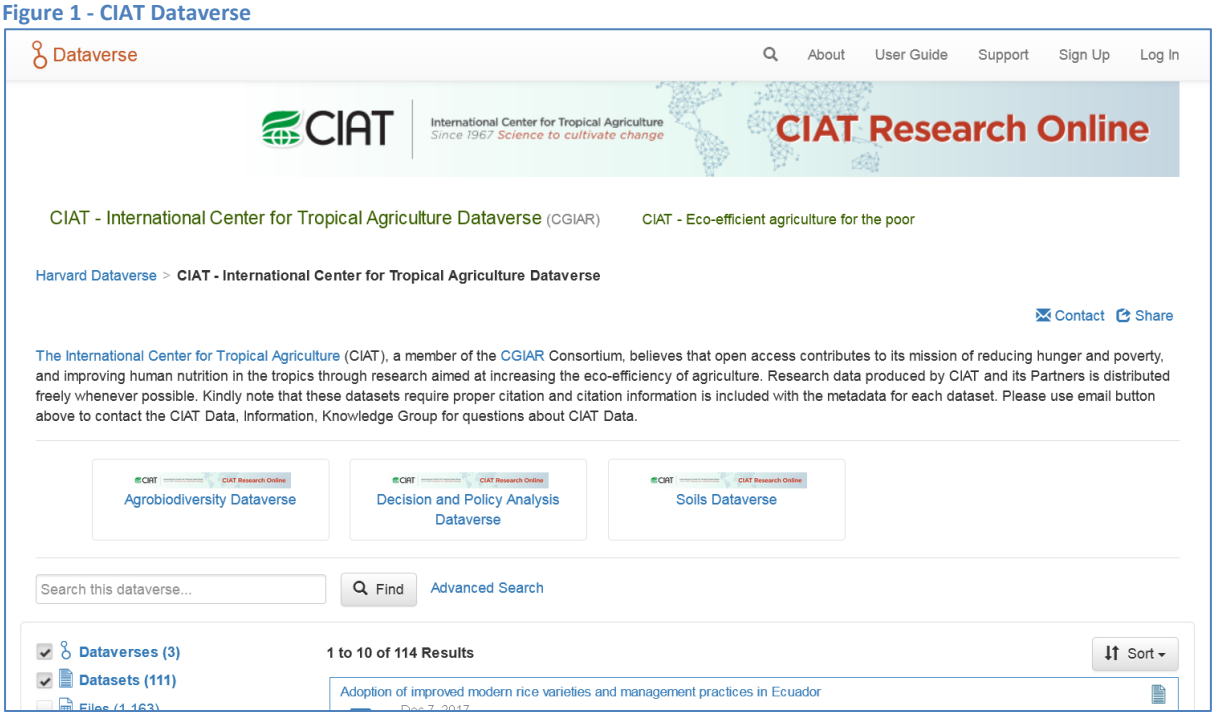
March 2018

Introduction

Dataverse is an open source web application for sharing, preserving, exploring and analysing research data. It was created by the Institute for Quantitative Social Science at Harvard University. Although aimed originally at social science data, a Dataverse can be created for research data on any subject. The website for the Dataverse Project is https://dataverse.org/

The image below shows the CIAT Dataverse which can be found at the following URL: https://dataverse.harvard.edu/dataverse/CIAT

Figure 1 - CIAT Dataverse



Structure of a Dataverse

A Dataverse is effectively a container for "datasets" where each "dataset" comprises research data, code, documentation and metadata. A Dataverse can also contain other Dataverses, i.e. Dataverses can be nested. For example, in Figure 1 you can see that the CIAT Dataverse includes three nested Dataverses: Agrobiodiversity Dataverse, Decision and Policy Analysis Dataverse and Soils Dataverse.

Citation

One thing that often stops researchers sharing their data is the belief that they won't get the credit for the work they have done. With Dataverse a unique citation is automatically generated when a dataset is created. For example, the unique citation for the CCAFS Household Baseline Survey in the CCAFS Dataverse is

> CCAFS, 2015, "CCAFS Household Baseline Survey 2010-2012", doi:10.7910/DVN/IUJQZV, Harvard Dataverse, V2, UNF:6:h/b2JxlvusEKXLXeRTwC7Q==

For further information about data citation please see the Data Citation page on the Dataverse website at http://best-practices.dataverse.org/data-citation/

Hosting

There are two options for hosting your Dataverse. The easiest is to have your Dataverse hosted at Harvard. This is a free service to all researchers. The infrastructure at Harvard is very good and great support is available to help you set up your Dataverse – you can also include your own branding in the header including working links so that your Dataverse has the look and feel of your own website. For example, the CIAT Dataverse includes a link in the header to take you directly to the CIAT website.

The other option is self-hosting and you can download and install the relevant software. This gives you more administrative control, but you would need an IT expert to install and manage the site including upgrading, taking backups, etc. You would also need good server infrastructure for hosting the application.

We would generally recommend the first option – i.e. having the Dataverse hosted at Harvard. To find out more about the two options see the online guides available on the Dataverse site. The User Guide is for those who wish to host their Dataverse at Harvard and the Installation Guide is for those who want to host their own Dataverse. The following link will take you to the Guides: http://guides.dataverse.org/en/latest/

Permissions

When a dataset is released, the default is for public access. However, you can choose to restrict the entire dataset giving access to named users only. Alternatively, you can restrict individual files within a dataset even if the dataset itself has public access.

Metadata

When you create a dataset, you will need to add metadata (formerly known as the Study Catalogue). The table below lists and describes the metadata elements for a dataset in Dataverse. Elements with a * before the name are compulsory elements. We recommend you use this list as a template for your own datasets and work on completing this list throughout the project. From experience we can tell you that finding this information at the end of a project is very difficult as those who know are likely to have moved on to other activities.

Element of Metadata	**Description**
**Title*	Full title by which the dataset is known
Subtitle	A secondary title used to amplify or state certain limitations on the main title
Alternative title	A title by which the work is commonly referred, or an abbreviation of the title
Alternative URL	A URL where the dataset can be viewed, such as a personal or project website
Other ID	Another unique identifier that identifies this dataset (e.g. producer's or another repository's number)

Element of Metadata	**Description**
**Author*	The person(s), corporate body(ies), or agency(ies) responsible for creating the work
**Contact*	The contact(s) for this dataset
**Description*	A summary describing the purpose, nature, and scope of the dataset
**Subject*	Domain-specific Subject Categories that are topically relevant to the Dataset. This is multiple response and you should select all relevant options from the list which is: • Agricultural Sciences • Other • Engineering • Business and Management • Computer and Information Science • Earth and Environmental Sciences • Physics • Chemistry • Law • Medicine, Health and Life Sciences • Arts and Humanities • Social Sciences • Astronomy and Astrophysics • Mathematical Sciences
Keyword	Key terms that describe important aspects of the Dataset
Topic Classification	The classification field indicates the broad important topic(s) and subjects that the data cover.
Related Publication	Publications that use the data from this dataset
Notes	Additional important information about the dataset
Language	Language of the dataset – select from the list
Producer	Person or organisation with the financial or administrative responsibility over this dataset
Production Date	Date when the data collection or other materials were produced (not distributed, published or archived)
Production place	The location where the data collection and any other related materials were produced
Contributor	The organisation or person responsible for either collecting, managing, or otherwise contributing in some form to the development of the resource
Grant Information	Grant information including grant agency, grant number, etc.
Distributor	The organisation designated by the author or producer to generate copies of the work including any necessary editions or revisions
Distribution Date	Date that the work was made available for distribution/presentation
Depositor	The person or the name of the organisation that deposited this dataset to the repository
Deposit date	Date that the dataset was deposited into the repository

Element of Metadata	**Description**
Time Period covered	Time period to which the data refer. This item reflects the time period covered by the data, not the dates of coding or making documents machine-readable or the dates the data were collected. Also known as the span.
Date of collection	Contains the date(s) when the data were collected
Kind of Data	Type of data included in the file: survey data, census/enumeration data, aggregate data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data/ratings, process-produced data or other
Series	Information about the dataset series
Software	Information about the software used to generate the dataset
Related Material	Any material related to this dataset
Related datasets	Any datasets that are related to this dataset such as previous research on this subject
Other references	Any references that would serve as background or supporting material to this dataset
Data sources	List of books, articles, serials or machine-readable data files that served as the sources of the data collection
Origin of Sources	For historical materials, information about the origin of the sources and the rules followed in establishing the sources should be specified
Characteristic of Sources Noted	Assessment of characteristics and source material
Documentation and Access to Sources	Level of documentation of the original sources

Summary

Dataverse is primarily an archiving facility. However, you can create a Dataverse and start populating it early on in your project, gradually building the archive as each stage of the project is completed. Datasets are only made public once they are released, so you can continue building your archive until you are ready to release it. If you have a Data and Document store for your project, then creating your archive should be relatively straight-forward as you can keep the same structure.

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes the following videos about Dataverse:

Introduction to Dataverse:
https://www.youtube.com/watch?v=EGYuj1JM1Qc&index=12&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj

Creating a Dataverse:
https://www.youtube.com/watch?v=9dMtCvCpZNM&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=13

CCAFS Dataverse:
https://www.youtube.com/watch?v=tr33h7TzFeY&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj&index=14